Machine Learning et nouvelles sources de données pour le scoring de crédit

Christophe Hurlin, Christophe Pérignon

Machine Learning et nouvelles sources de données pour le scoring de crédit*

Christophe Hurlin† Christophe Pérignon‡

16 décembre 2019

Résumé

Dans cet article, nous proposons une réflexion sur l'apport des techniques d'apprentissage automatique (Machine Learning) et des nouvelles sources de données (New Data) pour la modélisation du risque de crédit. Le scoring de crédit fut historiquement l'un des premiers champs d'application des techniques de Machine Learning. Aujourd'hui, ces techniques permettent d'exploiter de « nouvelles » données rendues disponibles par la digitalisation de la relation clientèle et les réseaux sociaux. La conjonction de l'émergence de nouvelles méthodologies et de nouvelles données a ainsi modifié de façon structurelle l'industrie du crédit et favorisé l'émergence de nouveaux acteurs. Premièrement, nous analysons l'apport des algorithmes de Machine Learning à ensemble d'information constant. Nous montrons qu'il existe des gains de productivité liés à ces nouvelles approches mais que les gains de prévision du risque de crédit restent en revanche modestes. Deuxièmement, nous évaluons l'apport de cette « datadiversité », que ces nouvelles données soient exploitées ou non par des techniques de Machine Learning. Il s'avère que certaines de ces données permettent de révéler des signaux faibles qui améliorent sensiblement la qualité de l'évaluation de la solvabilité des emprunteurs. Au niveau microéconomique, ces nouvelles approches favorisent l'inclusion financière et l'accès au crédit des emprunteurs les plus fragiles. Cependant, le Machine Learning appliqué à ces données peut aussi conduire à des biais et à des phénomènes de discrimination.

JEL classification : G21, G29, C10, C38, C55.

*Nous remercions Sébastien Saurin et Elisa Korn pour leur assistance et Jean-Paul Pollin pour ses commentaires et ses encouragements. Nous remercions également les participants à la table ronde « Pourquoi et comment les nouvelles technologies vont-elles bouleverser le secteur financier ? » de l'édition 2019 des Rendez-vous de l'Histoire (Blois). Ce travail a bénéficié du soutien financier de la Chaire ACPR « Régulation et Risque Systémique » et des programmes ANR MultiRisk (ANR-16-CE26-0015-01) et F-STAR (ANR-17-CE26-0007-01).

†Univ. Orléans, CNRS, LEO (FRE 2014), Rue de Blois, 45067 Orléans. E-mail : christophe.hurlin@univ-orleans.fr

‡HEC Paris, Département Finance, GREGHEC (UMR CNRS 2959), 1 Rue de la Libération, 78350 Jouy-en-Josas, E-mail : perignon@hec.fr

Introduction

Les décisions basées sur des algorithmes deviennent prépondérantes dans bon nombre de domaines tels que le diagnostic médical, la justice prédictive, la reconnaissance faciale, la détection de fraudes, la recherche d'emploi, ou l'accès à l'enseignement supérieur. Le monde de la finance n'échappe bien évidemment pas à cette révolution de la science des données. L'intelligence artificielle (IA) et les techniques d'apprentissage automatique (*Machine Learning* ou ML par la suite) sont particulièrement utiles en matière de connaissance client, d'allocation d'actifs, de détection de blanchiment et de transactions illégales, de gestion des risques, ou d'amélioration des processus internes (ACPR, 2018).

Dans cet article, nous nous concentrons sur une activité financière particulière, à savoir celle de l'évaluation du risque de crédit, et sur une technologie particulière, à savoir celle du ML. L'évaluation du risque de crédit est bien entendu une activité centrale pour tout établissement de crédit. Elle repose généralement à la fois sur une approche qualitative (expertise métier, relation clientèle) et une approche modélisée, au travers des modèles de risque. Ces modèles statistiques couvrent différentes dimensions du risque de crédit[1], mais dans la suite de cet article nous nous concentrerons sur les modèles de scoring. Rappelons qu'un modèle de scoring de crédit a pour finalité de *prévoir* le risque de non-remboursement du prêt, appelé risque de crédit ou de défaut[2]. Il peut être utilisé pour décider de l'octroi de nouveaux prêts ou pour évaluer le niveau de risque du portefeuille de crédits en production. Dans ce dernier cas, les modèles de scoring peuvent être utilisés dans un contexte réglementaire que ce soit pour déterminer les provisionnements sur les pertes (IFRS 9) ou le montant des fonds propres exigés (Bâle III). Quelle que soit leur finalité, ces modèles permettent d'obtenir *in fine* une *classification* des crédits en fonction de leur niveau de risque. Cette classification se présente généralement sous la forme de classes de risques homogènes.

Même si par nature, le scoring de crédit est un sujet technique, il revêt une importance capitale au niveau économique et sociétal. En effet, il conditionne l'allocation du crédit entre les agents économiques : quels ménages vont pouvoir accéder à la propriété, quelles entreprises vont pouvoir financer leurs programmes d'investissements, quelles sociétés vont devoir déposer leur bilan et parmi celles-ci combien seront liquidées, etc. Dès lors, les modèles de scoring de crédit ont des implications majeures en termes de stabilité financière (provisions, capital réglementaire des banques), d'inclusion financière, d'emploi, et de croissance économique.

[1] Les techniques de ML sont aussi utilisées pour la modélisation de la *Loss Given Default*, c'est-à-dire du taux de perte en cas de défaut (voir Loterman et al. (2012) pour une synthèse) ou du *Credit Conversion Factor*.

[2] La définition du défaut peut être réglementaire (voir la définition du défaut proposée par l'EBA (2016) dans le cadre IRB qui entrera en vigueur au 1er janvier 2021) ou interne, par exemple dans le cas d'un score d'octroi.

Le scoring de crédit fut historiquement l'un des premiers champs d'application des techniques de ML. Nous définissons dans cet article le ML comme un « ensemble d'algorithmes destinés à résoudre des problèmes et dont la performance s'améliore avec l'expérience et les données sans intervention humaine a posteriori » (ACPR 2018). Ainsi, un algorithme de ML[3] est un programme informatique qui permet de construire, et surtout d'améliorer de façon autonome un modèle de régression ou de classification. Dans le contexte du risque de crédit, nous considérerons principalement des modèles de classification. Un algorithme de classification vise à établir une fonction de lien[4] (un modèle) entre une variable cible Y (de type binaire dans le cas du scoring de crédit, par exemple défaut ou non défaut) et un ensemble de prédicteurs ou caractéristiques X. Cette fonction de lien est révélée à partir d'un échantillon d'apprentissage. La méthode de classification est dite supervisée lorsque la variable cible Y est observée sur l'échantillon d'apprentissage. Une fois le modèle entrainé sur l'échantillon d'apprentissage, il est ensuite utilisé pour réaliser une prévision (classification) de la variable Y sur un échantillon test à partir des observations des prédicteurs X, la distribution conditionnelle de Y sachant X étant supposée être la même dans les deux échantillons.

Il existe de très nombreuses méthodes de ML (voir Varian (2014) ou Mullainathan et Spiess (2017) pour une typologie), mais de façon générale, on peut distinguer deux grandes familles. Premièrement, les méthodes de classification supervisée dites « individuelles » visent à partitionner l'espace des prédicteurs afin de prévoir l'événement. Les plus utilisées dans le domaine du risque de crédit sont les arbres de classification, les machines à vecteurs de support (SVM), et les réseaux neuronaux. Deuxièmement, les méthodes d'ensemble combinent des prédictions issues d'un ensemble de plusieurs modèles de base en utilisant une règle de vote pour aboutir à la classification finale. Les méthodes d'ensemble les plus couramment utilisées dans le domaine sont le *Bagging*, les forêts aléatoires (*Random Forests*), ou le *Boosting*. Le *Bagging* consiste à entrainer un ensemble d'arbres de classification sur des sous-échantillons d'individus tirés au hasard. Les *Random Forests* reposent sur le même principe avec en outre un tirage aléatoire des prédicteurs à chaque branche de l'arbre de classification. Dans les deux cas, la prévision finale est alors construite en agrégeant les prévisions obtenues sur tous les arbres. Le *Boosting*[5] consiste à entrainer de façon itérative un modèle de base de façon à

[3] Les lecteurs intéressés par la distinction entre ML, IA, économétrie et statistique pourront se reporter aux travaux de Varian (2014), Mullainathan et Spiess (2017), Charpentier, Flachaire et Ly (2018), Athey et Imbens (2019), ou Athey (2019).

[4] Pour certains algorithmes, la fonction de lien relie directement la variable cible Y (défaut, non défaut) aux prédicteurs. Pour d'autres méthodes (comme la régression logistique), l'output du modèle correspond à la probabilité conditionnelle $\Pr(|Y = 1\ X = x)$, à partir de laquelle il est possible dans un second temps de prévoir le défaut en comparant cette probabilité à un seuil donné, typiquement 50%.

[5] Certaines méthodes de *Boosting* comme l'*Extreme Gradient Boosting* (XGBoost) figurent aujourd'hui parmi les plus utilisées dans la plupart des compétitions de ML.

réduire les erreurs de prévision à chaque étape.

Un premier avantage du ML est sa capacité à sélectionner de manière extrêmement flexible la forme fonctionnelle du lien entre la variable cible Y et les prédicteurs X. En clair, l'analyste fournit une liste de prédicteurs potentiellement mobilisables, puis l'algorithme évalue de nombreux modèles alternatifs et sélectionne l'un d'eux afin de maximiser un critère donné. Ce faisant, l'algorithme de ML sélectionne de façon autonome les variables explicatives qui seront au final introduites dans la spécification du modèle. En cela, le ML s'apparente aux méthodes automatiques de sélection de modèles économétriques (*stepwise*, *forward*, *Autometrics*, etc.). Cependant, l'avantage du ML va bien au-delà de cette sélection automatique puisque ces algorithmes peuvent créer de « nouveaux » prédicteurs en combinant et transformant les prédicteurs initiaux. Cette transformation de l'espace de représentation des données rend ces méthodes particulièrement efficaces pour détecter de façon autonome des non-linéarités et des interactions entre les prédicteurs. Une question est de savoir si la prise en compte de ces interactions permet d'obtenir des gains prédictifs, c'est-à-dire une meilleure évaluation des risques de crédit. Quoiqu'il en soit ces techniques génèrent d'important gains de productivité dès lors qu'elles rendent caduques un certain nombre de prétraitements sur les données, visant notamment à capter ces non-linéarités.

Mais l'avantage essentiel du ML est qu'il permet l'utilisation de nouvelles données (*New Data*) susceptibles de mieux rendre compte du risque de crédit. Cette « datadiversité » est rendue possible par les nouvelles pratiques des clients, la digitalisation de la relation clientèle ou l'accessibilité à de nouvelles sources d'information (réseaux sociaux). Le caractère novateur s'apprécie ici par rapport aux données clientèles habituellement utilisées dans les modèles de score telles que les historiques de paiements, le revenu, etc. La question centrale est de savoir si cette « datadiversité » permet in fine un accès au crédit d'individus ou d'entreprises jusque-là considérés comme trop risqués dans les bases de données traditionnelles. Enfin, ces nouvelles données peuvent être considérées comme des *Big Data* si le nombre de prédicteurs observés pour chaque individu est très important, si l'on observe des prédicteurs pour un grand nombre d'individus, ou si les données sont massives et occupent énormément de place en mémoire. Quelles soient massives ou non, ces nouvelles données peuvent être inclues directement dans les modèles de scoring des établissements de crédit, mais elles sont le plus souvent collectées et traitées par des prestataires externes. Ces Fintechs, quand elles ne distribuent pas elles-mêmes du crédit, proposent alors des scores sur les clients qui sont introduits en tant que variable explicative afin d'enrichir les modèles de scoring des établissements de crédit. Mais une telle externalisation de la connaissance du risque de crédit ne va pas sans poser problème, que ce soit en termes de pouvoir de marché, de responsabilité juridique au regard de la légalité des

données collectées et utilisées (par exemple vis-à-vis du RGDP en Europe), ou de possibles biais. Plus généralement, le développement du ML dans le contexte de la modélisation du risque de crédit n'est pas sans risques. Ces risques concernent à la fois les algorithmes eux-mêmes, leur mise en œuvre, et l'organisation même du secteur de la distribution du crédit. Ces enjeux appellent une réflexion normative sur la régulation de ces nouvelles approches.

Dans cet article, nous proposons une première analyse critique et prospective de l'application des techniques de ML dans le contexte de la modélisation du risque de crédit en traitant les questions suivantes. Quels sont réellement les apports du ML pour la connaissance du risque de crédit ? Que peut-on attendre des algorithmes eux-mêmes lorsqu'on les applique à périmètre constant sur des données traditionnelles ? Quelles en sont les limites ? Quel est l'apport des nouvelles données vis-à-vis de la connaissance du risque de crédit et de l'inclusion financière ? Quels sont les risques associés à ces nouvelles données, couplées à l'utilisation du ML ?

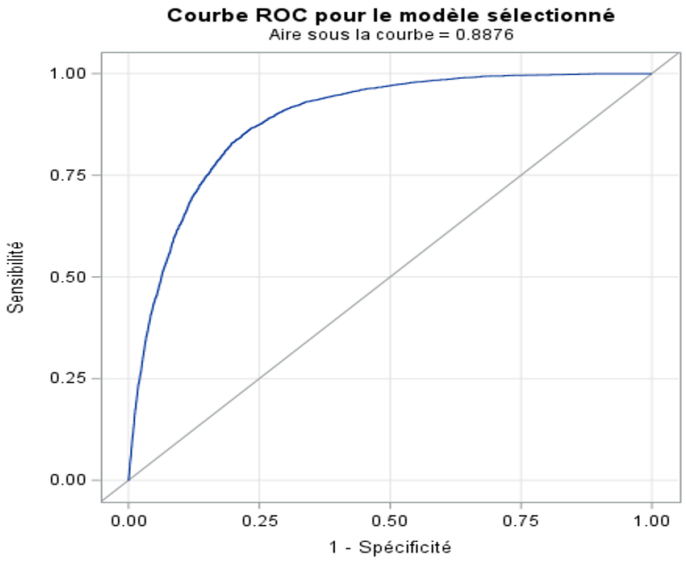
1 L'apport du ML au scoring de crédit

1.1 Les gains prédictifs

Tout modèle de scoring est construit à partir d'un échantillon de n crédits pour lesquels on observe la survenue ou la non-survenue d'un défaut, représenté par une variable dichotomique Y. Pour chaque individu de cet échantillon, on dispose également d'un ensemble de variables explicatives ou prédicteurs, qui correspondent par exemple à des informations sur la nature du contrat et/ou sur l'emprunteur. Cette base de données est alors décomposée en deux sous-échantillons : un échantillon d'apprentissage sur lequel le modèle est sélectionné, étalonné et éventuellement estimé, et un échantillon de test sur lequel on évalue la performance prédictive *out-of-sample* du modèle de risque. Pour les algorithmes de ML, l'échantillon d'apprentissage est généralement lui-même décomposé en deux sous-échantillons : un échantillon sur lequel l'algorithme de classification est entrainé et un échantillon de validation qui permet de déterminer la valeur des hyperparamètres (ou paramètres de *tuning*) associés à la méthode de classification et de contrôler ainsi le phénomène de sur-apprentissage. L'idée est alors de déterminer la valeur des hyperparamètres qui maximise une mesure de performance calculée sur un échantillon (l'échantillon de validation) différent de celui sur lequel l'algorithme est entrainé (l'échantillon d'apprentissage). Ainsi, cette approche réduit le risque de surajustement qui peut être induit par la fixation de valeurs « optimales » pour les hyperparamètres qui permettraient de reproduire quasi-parfaitement la classification sur l'échantillon d'apprentissage, mais qui conduirait au final à de très mauvaises performances de classification *out-of-sample*.

Cette démarche peut être généralisée à une approche de validation croisée de type *k-fold*[6] appliquée sur l'ensemble de l'échantillon d'apprentissage.

FIG. 1 – Exemple de courbe ROC



Une fois le modèle calibré (pour les algorithmes de ML) ou estimé (pour les approches paramétriques usuelles), il est appliqué à l'échantillon de test. Suivant les modèles, on obtient alors pour chaque individu de l'échantillon de test soit une estimation de la probabilité conditionnelle de survenue de l'événement de défaut, comme par exemple dans le cas d'une régression logistique, soit directement une prévision de cet événement représentée sous la forme d'une variable dichotomique $\widehat{Y}$, comme par exemple dans le cas d'un arbre de classification. Dans le cas, où les modèles produisent des probabilités estimées, on se ramène alors à une prévision sur l'événement en comparant la probabilité à un seuil c, typiquement 50%. Si la probabilité excède ce seuil, on prévoit la survenue de l'événement, c'est-à-dire $\widehat{Y}(c) = 1$. Pour un seuil donné, on peut alors construire une matrice de confusion recensant les occurrences de deux types d'erreurs de classification commises sur l'échantillon de test. Les faux positifs correspondent aux individus pour lesquels le modèle avait prévu un défaut ($\widehat{Y}(c) = 1$) mais pour lesquels aucun défaut n'a été observé ex-post ($Y = 0$). A l'inverse, les faux négatifs correspondent aux individus pour lesquels le modèle n'avait pas prévu de défaut, et pour lesquels

[6] L'idée de la méthode du *k-fold* consiste à diviser l'échantillon d'apprentissage initial en k segments, puis de sélectionner un des k segments comme échantillon de validation et d'entrainer l'algorithme sur les k-1 autres segments. En répétant l'opération k fois, on peut alors construire une erreur de prédiction et déterminer la valeur des hyper-paramètres qui minimise cette erreur tout en évitant les problèmes de surajustement.

un défaut a été observé. Ces erreurs peuvent être exprimées sous forme de ratios, tels que la spécificité et la sensibilité. La sensibilité correspond à la probabilité de prévoir le défaut dans la population des défauts, tandis que la spécificité est la probabilité de prédire un non défaut dans la population des non défauts. A partir de ces éléments, on peut alors construire la courbe ROC (*Receiver Operating Characteristic*), dont les éléments correspondent à la sensibilité (axe des abscisses) et la spécificité (axe des ordonnées) obtenues pour des valeurs du seuil c variant de 0 à 1 (voir Figure 1). L'intérêt de la courbe ROC est de permettre d'évaluer la capacité prédictive du modèle de classification indépendamment du choix du seuil[7]. Un des critères usuels est alors donné par l'aire sous cette courbe : l'AUC (*Area Under the Curve*). Plus cette aire est grande, plus la courbe ROC s'écarte de celle d'un classificateur aléatoire (AUC=1/2) et se rapproche de celle d'un classificateur parfait (AUC=1) qui ne comment aucune erreur de type 1 et 2.

Dès le milieu des années 80, de nombreuses études académiques ont cherché à évaluer les gains de performance prédictive des méthodes de ML par rapport à la régression logistique. Trente ans plus tard, le diagnostic est relativement mitigé. Makowski (1985), Coffman (1986), Srinivasan et Kim (1987) et Carter et Catlett (1987) furent parmi les premiers[8] à appliquer des arbres de classification pour le scoring de crédit dans le but de capter les interactions entre les prédicteurs. Les réseaux de neurones artificiels furent également très rapidement appliqués, principalement sur des problèmes de scoring d'établissements bancaires (Tam et Kiang, 1992) ou d'entreprises (Altman, Marco et Varetto, 1994). Cette dernière étude conclut de façon mitigée en pointant notamment l'aspect boite noire des réseaux de neurones, le poids parfois illogique accordé à certains prédicteurs, et les problèmes de surajustement. Desai, Crook et Overstreet (1996) comparent différentes sortes de réseaux de neurones aux techniques standards telles que la régression logistique et l'analyse linéaire discriminante, sur une base de crédits aux particuliers. Ils montrent que les réseaux de neurones offrent de très bonnes performances prédictives dès lors que l'on s'intéresse au pourcentage de mauvais crédits correctement classifiés. En revanche, les performances prédictives des réseaux de neurones sont similaires à celles de la régression logistique en ce qui concerne le pourcentage de bons et de mauvais crédits correctement identifiés.

De façon générale, il apparait que les classifieurs individuels de ML ne permettent pas d'améliorer significativement les performances prédictives de la régression logistique. Ces ré-

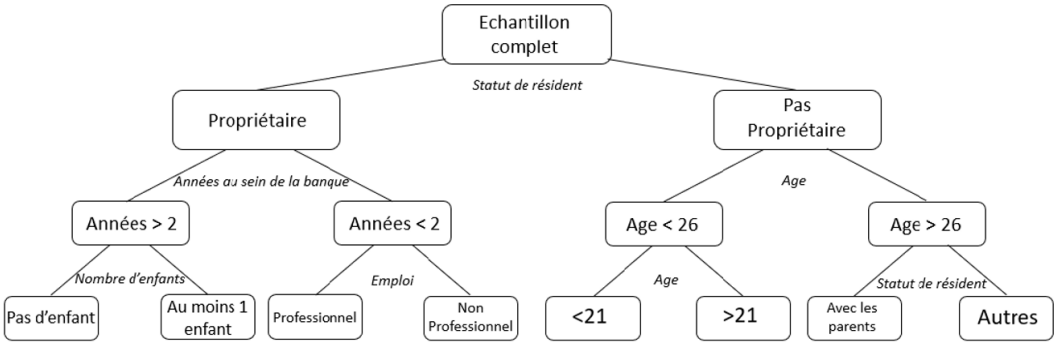
[7] Il existe de nombreux critères alternatifs pour juger de la performance prédictive d'un modèle de scoring, tels que le pourcentage de classification correct, l'indice de Gini partiel, le score de Brier, etc. Pour une discussion complète sur ces critères d'évaluation voir Candelon, Dumitrescu et Hurlin (2012) ou Lessmann et al. (2015).

[8] Notons que l'algorithme CART (*Classification And Regression Trees*) n'a été publié par Leo Breiman et se co-auteurs qu'en 1984, soit seulement un avant les premières applications en scoring de crédit.

sultats sont confirmés par Thomas (2000) qui propose la première synthèse de la littérature concernant les modèles de scoring incluant des techniques de ML. L'auteur reporte le pourcentage de classification correcte (PCC) de six méthodes (arbres de classification, réseaux de neurones, régression logistique, régression linéaire, etc.) issues de cinq études. Il montre qu'aucune méthode ne domine les autres, mais surtout que les différences entre les PCC de ces différentes méthodes sont très faibles. Ces résultats sont confirmés par l'étude comparative de Baesens et al. (2003) qui propose une analyse systématique de 17 algorithmes de classification à partir de huit bases de données de prêts fournies par des banques internationales. Pour la plupart des bases considérées, les machines à vecteurs de support (SVM) ou les réseaux de neurones offrent de très bonnes performances prédictives avec des AUC variant de 66% à 91%. Mais les auteurs montrent aussi que pour la plupart des bases, les différences entre les AUC de la meilleure méthode de ML et celle de la régression logistique sont inférieures à 2%.

Comment expliquer des performances prédictives si peu contrastées ? Le principal avantage de ces algorithmes de ML par rapport aux approches paramétriques standards réside dans leur capacité à révéler automatiquement les interactions entre les prédicteurs et des non linéarités (effets de seuils). Considérons l'exemple d'un arbre de classification tel que celui représenté sur la Figure 2. La classification d'un crédit en mauvais ou bon risque prend la forme d'un arbre qui se sépare en deux à chaque nœud. La valeur d'un prédicteur (par exemple le statut résidentiel) détermine si la branche de droite (non propriétaire) ou celle de gauche (propriétaire) doit être considérée pour la suite de l'algorithme. A la fin de l'algorithme, lorsque le dernier nœud est atteint, le crédit est affecté à une feuille et à une prévision (0 ou 1). Cette prévision correspond à la classe majoritaire (0 ou 1) des observations appartenant à ce nœud. Par exemple, imaginons que sur les 1000 crédits de l'échantillon initial, 120 crédits aient été attribués à des clients (i) propriétaires, (ii) avec plus de deux ans d'ancienneté dans la banque, (iii) et sans enfant. Si parmi ces 120 crédits qui sont affectés à la feuille gauche de l'arbre la fréquence de défaut est faible, par exemple de 14%, alors on prédit l'absence de défaut pour tous les crédits présentant ces caractéristiques. Au final, tout se passe comme si l'on considérerait un modèle de régression dans lequel seraient introduites des variables explicatives binaires définies par des produits (ou interactions) des prédicteurs initiaux. Par exemple, l'arbre de la Figure 2 revient au final à construire une première variable explicative valant 1 si le client est propriétaire, s'il a plus de 2 ans d'ancienneté dans sa banque et s'il n'a pas d'enfant. Ainsi, les arbres de classification permettent de capter des interactions entre les prédicteurs initiaux et des effets non linéaires, typiquement des effets de seuil dans ce cas, qu'il aurait été difficile de repérer dans une approche paramétrique standard sans évaluer un nombre infiniment grand de combinaison et de seuils. De façon générale, on retrouve une idée similaire dans nombre

FIG. 2 – Exemple d'arbre de classification. Source : d'après Thomas (2000)



d'algorithmes de ML (réseaux de neurones, machines à vecteurs de support, etc.) au travers de la notion de transformation de l'espace de représentation des données. La question est alors de savoir si l'événement modélisé présente ce type de non-linéarités. Or, la conclusion de l'étude de Baesens et al. (2003) est que le scoring de crédit est un champ d'application dans lequel il y a au final trop peu de non-linéarités dans les données usuelles pour que les gains de performances prédictives du ML soient significatifs.

C'est finalement l'utilisation des premières méthodes d'ensemble dans les années 2000 qui va permettre d'obtenir des gains prédictifs significatifs. L'intuition de ces approches est de combiner différents modèles de classification élémentaires susceptibles d'apporter des informations complémentaires. On retrouve ainsi l'idée d'une combinaison automatique de prévisions ou de modèles. Douze ans après l'étude de Bart Baesens, Lessmann et al. (2015) proposent une nouvelle analyse comparative en utilisant d'autres critères d'évaluation (score de Brier, H-*measure*, etc.) et les algorithmes de ML les plus récents, y compris des méthodes d'ensemble basées sur le principe du *Bagging* ou sur celui du *Boosting*. Au final, leur étude porte sur 41 algorithmes de classification appliqués sur 8 bases de données de crédits aux particuliers. Leur conclusion est beaucoup plus en faveur du ML : plusieurs méthodes d'ensemble prévoient le risque significativement mieux que la régression logistique. Par exemple, les *Random Forests* dominent systématiquement les classifieurs individuels que ces derniers soient paramétriques (régression logistique) ou de type ML (arbres, réseaux de neurones, SVM, etc.). Les meilleures performances sont obtenues pour des méthodes d'ensemble hétérogènes comme la méthode *Weighted Average Ensemble*. Le second enseignement de cette étude est que les gains de performance prédictive liés au ML ont tendance à plafonner. Les raffinements méthodologiques des algorithmes de ML n'améliorent pas nécessairement les performances des modèles de score. Par exemple, les AUC des *Rotation Forests* ne diffèrent pas significativement de ceux des

Random Forests.

La question centrale qui reste en suspens est de savoir pourquoi certains algorithmes de ML présentent de bonnes performances prédictives. La réponse n'est pas évidente et aucune règle ne semble se dégager. A ce jour, aucune recherche n'a encore permis d'expliquer la performance de ces classifieurs en fonction de leurs caractéristiques et des caractéristiques des bases des données.

1.2 ML et gains de productivité

Au-delà de la question de la performance prédictive, les méthodes de ML présentent un avantage indéniable par rapport aux approches paramétriques usuelles de scoring, puisqu'elles permettent des gains de productivité importants. En particulier, les algorithmes de ML permettent de réduire le temps consacré aux étapes de gestion et de prétraitement des données, avant l'étape de modélisation au sens strict. Bien évidemment, cela ne signifie pas que le ML permette de se dispenser du travail de construction et de contrôle de la qualité des données, qui reste absolument nécessaire.

Pour bien comprendre ce point, revenons à la démarche traditionnelle d'un statisticien en charge de la construction d'un modèle de scoring de crédit au sein de la direction des risques d'une grande banque. La première étape de son travail consiste à appliquer différents traitements aux données d'apprentissage. Parmi ceux-ci figure tout d'abord le traitement des valeurs manquantes ou aberrantes, qui nécessite la mise en œuvre de procédures de détection, d'imputation et d'exclusion. Les autres traitements concernent généralement le regroupement de modalités des variables explicatives discrètes et la discrétisation des variables continues. Pour chacune des variables qualitatives, les modalités sont regroupées de façon à réduire le nombre de classes et à maximiser le pouvoir discriminant de la variable. Toutes les variables explicatives continues sont quant à elles discrétisées. Il s'agit ici d'une part de capter de potentiels effets non linéaires et d'autre part de réduire l'influence des valeurs extrêmes ou valeurs aberrantes non corrigées. Le nombre de classes et les seuils de discrétisation sont déterminés par des algorithmes itératifs construits dans l'objectif de maximiser une mesure d'association de type V de Cramer ou statistique du khi-deux, entre la variable cible (le défaut) et la variable explicative. La seconde étape consiste en l'analyse des corrélations entre les prédicteurs afin de vérifier que ces variables ne sont pas trop corrélées entre elles. En fonction de ces corrélations, l'expert décide alors de supprimer certaines variables redondantes suivant un principe de parcimonie. La troisième étape est celle de la sélection des variables explicatives du modèle de score. Dans le cadre d'un modèle de score donné (par exemple une régression logistique), on sélectionne parmi toutes les variables retraitées celles qui permettent

de prévoir au mieux le défaut. Suivant le nombre de variables à disposition, cette sélection peut être réalisée soit manuellement, soit grâce à des approches automatiques telles que le *stepwise*. La sélection automatique est souvent complétée par une expertise métier et une analyse plus fine du modèle (effets marginaux, *odds ratios*).

A l'inverse, l'utilisation d'un arbre de classification ou d'algorithmes basés sur des arbres comme les *Random Forests*, rend caduque le travail de discrétisation des variables continues et de regroupement des modalités. Par essence, ces techniques déterminent de façon autonome les discrétisations optimales et les regroupements de modalités. L'analyse des corrélations entre les prédicteurs est moins cruciale dans le sens où la plupart des algorithmes de ML peuvent intégrer des prédicteurs fortement corrélés. Les méthodes de régression pénalisées telles que le *Lasso* ou le *Ridge* permettent précisément de sélectionner les variables pertinentes et de pallier au problème de la multi-colinéarité. De façon plus générale, l'avantage des algorithmes de ML est précisément d'utiliser les données pour déterminer la forme fonctionnelle optimale du modèle au sens d'un certain critère. Cela rend donc caduque l'étape de sélection des variables explicatives du modèle de score.

Ces gains de productivité associés au ML[9] sont aujourd'hui mis en avant dans l'industrie financière. Grennepois, Alvirescu, et Bombail (2018) soulignent le fait que les performances prédictives des *Random Forests* sont généralement robustes à la non-imputation des valeurs manquantes, à la présence de fortes corrélations entre certaines variables explicatives, au non regroupement des modalités des variables discrètes, et à la non-discrétisation des variables continues. Cette robustesse permet donc potentiellement de limiter les étapes de prétraitement sur les données. Au-delà de gains de productivité, le fait de limiter les prétraitements sur les données peut en outre réduire les éventuels biais de modélisation puisque qu'au final, le ML laisse parler les données brutes. L'utilisation du ML permet ainsi une automatisation accrue des processus d'octroi de crédit, y compris dans la phase de construction et de révision des modèles de risque. Considérant des données sur la durée de traitement de demandes de prêts hypothécaires aux Etats-Unis, Fuster et al. (2018a) montrent que les Fintechs traitent les demandes de prêt environ 20% plus rapidement que les autres prêteurs, et cela sans détérioration notable de la qualité de la sélection des dossiers.

1.3 Facilité de mise en œuvre du ML

Les gains de productivité et de performances prédictives offerts par le ML sont d'autant plus accessibles que l'on assiste aujourd'hui à une véritable démocratisation de l'usage de

[9] Voir aussi Phaure et Sartre (2019) pour l'utilisation des techniques de classification non ou semi supervisées pour l'analyse des risques de concentration dans les portefeuilles de crédit.

ces algorithmes. Cet usage s'est grandement simplifié grâce au développement de procédures simplifiée et performantes. Dans le domaine de la modélisation du risque de crédit, les principaux logiciels utilisés dans l'industrie sont SAS, R, et Python. Chacun d'entre eux offre de nombreuses procédures, packages, ou environnements permettant de mettre en œuvre les principales techniques de ML. On peut citer entre autres les librairies *Scikit Learn* ou *Keras* pour Python, les packages *caret*, *mlr* ou *e1071* pour le logiciel R, le système *Visual Data Mining and Machine Learning* et les procédures *High Performance* pour SAS. Certains outils comme SAS Viya proposent une approche unifiée sur l'ensemble du cycle de vie des données, allant de l'analytique, au traitement par ML, jusqu'à la visualisation des résultats. De même, des bibliothèques logicielles open source permettent de mettre en œuvre des algorithmes plus spécifiques sous différents langages et environnements de programmation. On peut citer par exemple la bibliothèque XGBoost qui permet d'implémenter des techniques avancées de *Gradient Boosting* sous R, Python et Julia. Un point important est que la plupart de ces packages proposent des procédures automatiques de détermination des hyperparamètres, généralement par des techniques de validation croisée. Ainsi l'utilisateur, même novice, peut lancer en une ligne de code ou par un simple clic de souris, une chaine complète de traitements mettant en œuvre un algorithme de ML.

La limite de ces langages est cependant rapidement atteinte lorsque l'on a affaire à d'énormes jeux de données qui posent des problèmes de mémoire. Toutefois, de nouvelles technologies permettent aujourd'hui d'appliquer facilement ces procédures de ML sur des données massives. Ainsi, le système *Hadoop* est un système de stockage assuré par fichiers distribués qui autorise une parallélisation des traitements, via l'outil *MapReduce*. Ce dernier permet le déploiement de traitements massivement parallèles sur des clusters de serveurs. Enfin, grâce au développement de l'industrie du cloud, il est aujourd'hui relativement aisé de déployer une infrastructure permettant d'utiliser des outils de ML sur des données massives. Les grands acteurs du secteur offrent de plus en plus de solutions d'infrastructure avec des outils de ML intégrés, comme par exemple *Amazon SageMaker* ou *Google Cloud AutoML*. Ces solutions permettent aux développeurs d'entrainer et de déployer rapidement des modèles de ML sur des infrastructures adaptées. Tous ces moyens autorisent une large diffusion des techniques de ML dans l'industrie bancaire et dans les Fintech.

1.4 Les limites de l'approche par ML

Le problème de la non-interprétabilité des algorithmes de ML. Une régression logistique ou un modèle de régression linéaire sont des modèles transparents. Ils attribuent des pondérations bien définies à chacune des variables explicatives. Toutefois, cette transparence

a un coût : ces modèles économétriques imposent une forme de relation particulière entre la variable dépendante et les facteurs explicatifs. Bien évidemment, cette forme spécifique de relation peut ne pas correspondre aux données. A l'inverse les algorithmes de ML sont plus flexibles : la forme fonctionnelle de la relation étudiée n'est pas fixée ex-ante. Ces modèles permettent donc de détecter des nuances plus fines à condition de disposer de suffisamment de données pour former un modèle. Cependant, cette flexibilité a un coût : c'est celui de l'opacité. En effet, pour certaines méthodes de ML il est difficile, voire impossible, de savoir quelles sont les variables qui sont à la base des prévisions du modèle. Ces algorithmes s'apparentent dès lors à des « boites noires » qui associent à un ensemble de prédicteurs des prévisions sur la variable cible sans que l'on sache expliquer l'origine de ces prévisions. Ceci est particulièrement vrai pour les méthodes d'ensemble comme le *Bagging* ou le *Boosting* qui par ailleurs présentent souvent les meilleures performances prédictives. Bien évidemment, cette opacité soulève de forts enjeux éthiques, juridiques, mais aussi en matière de régulation financière lorsque ces modèles sont utilisés pour des décisions affectant la vie des individus ou des entreprises, comme par exemple l'octroi d'un crédit.

Un algorithme de ML est dit interprétable s'il est possible d'identifier les prédicteurs qui participent le plus à la règle de décision et d'en quantifier l'importance. Cela revient à la question du caractère intelligible ou non du fonctionnement de l'algorithme. Par exemple, un arbre de classification est généralement interprétable : il est possible de comprendre comment fonctionne la classification des crédits en fonction des prédicteurs. A l'inverse, une méthode d'agrégation d'arbres construits sur des échantillons Bootstrap, comme par exemple les *Random Forests*, est non interprétable. Une décision algorithmique est dite explicable s'il est possible de rendre compte explicitement d'une décision individuelle (par exemple la classification d'un crédit en catégorie risquée) à partir des caractéristiques associées cet individu. Pour une discussion générale sur ce sujet, voir Molnar (2019).

Lorsque les algorithmes de ML ne sont pas interprétables a priori, il existe aujourd'hui des méthodes, dites *Model-Agnostic Methods*, pour les rendre interprétables ex-post. L'idée générale consiste à faire de l'ingénierie inversée à partir des seuls inputs (prédicteurs) et outputs (prévisions) afin de révéler la règle de décision de l'algorithme. Une première approche est celle des graphiques de dépendances partielles (ou PDP pour *Partial Dependencies Plot*). Le but est de balayer sur un ensemble de valeurs possibles pour une variable explicative d'intérêt, de refaire tourner l'algorithme de ML en utilisant pour tous les individus de la base d'apprentissage ces valeurs successives (en laissant inchangées les valeurs des autres variables explicatives), et de calculer pour chaque valeur testée la moyenne des scores obtenus. Un graphique de PDP permet de représenter l'effet marginal d'une caractéristique sur le résultat

attendu d'un modèle de ML. On peut ainsi vérifier si la relation entre la cible et un prédicteur est linéaire, monotone ou plus complexe. Toutefois, la principale limite des PDP est l'hypothèse d'indépendance : on suppose que la caractéristique pour laquelle la dépendance partielle est calculée n'est pas corrélée avec d'autres caractéristiques.

Une approche alternative repose sur l'utilisation de modèles de substitution (*surrogate models*). L'idée est alors d'approximer la règle de décision non interprétable d'un algorithme de ML par un modèle de substitution plus simple et interprétable, considéré comme une approximation locale ou globale[10] de la vraie règle de décision. Un exemple d'approximation locale est la méthode LIME (*Local interpretable model-agnostic explanations*). L'intuition est simple : pour un individu de référence, on génère aléatoirement des variables explicatives d'individus proches de l'individu de référence, et l'on applique alors l'algorithme de ML sur ces données simulées. A partir des caractéristiques simulées et des prévisions qui en découlent (pondérées par rapport à leur distance à l'individu de référence), LIME forme un modèle interprétable. Il s'agit typiquement d'un arbre de décision ou d'une régression pénalisée de type *Lasso*. On peut alors représenter l'influence approximée de chacun des prédicteurs sur la décision d'un individu de référence. Ainsi, on peut expliquer pourquoi tel emprunteur a été classé en défaut et quelles sont les caractéristiques qui ont conduit à ce choix.

D'autres méthodes visent à expliquer les décisions au niveau individuel. On peut citer ici les valeurs de Shapley dont l'intuition est la suivante : une prévision pour un individu peut être assimilée à un « jeu » dans lequel chacune des caractéristiques de l'individu est un « joueur » et où la prévision est le paiement. La question est alors de savoir comment répartir équitablement le paiement entre les joueurs. Les valeurs de Shapley (du nom de Lloyd Shapley, lauréat du prix Nobel d'économie en 2012) sont fondées sur l'idée que cette répartition doit être déterminée par l'excédent généré par chaque joueur. Dans le contexte du ML, ce gain sera défini comme la différence entre la prévision pour l'individu de référence et la moyenne des prévisions pour les autres individus. Ainsi, la contribution de Shapley d'une variable correspond à la contribution marginale moyenne (pondérée) de cette caractéristique pour toutes les combinaisons possibles dans lesquelles cette variable aurait pu être ajoutée à l'ensemble des autres variables explicatives. Cette décomposition a pour avantage d'être fondée sur une théorie solide, mais elle peut s'avérer couteuse en temps de calcul en fonction du nombre de prédicteurs, d'observations, et de la complexité du modèle considéré. Bracke et al. (2019) proposent une approche englobante, appelée *Quantitative Input Influence* (QII), basée sur les valeurs de Shapley, mais à partir de laquelle les auteurs tirent une interprétation

[10] Un modèle de substitution est dit « global » lorsqu'il vise à approximer le fonctionnement du modèle complexe pour toutes les données d'entrée. Un modèle est dit « local » lorsqu'il vise à expliquer ce même fonctionnement uniquement pour une sélection particulière des individus de l'échantillon.

globale de l'influence des variables. Cette représentation graphique, similaire dans l'esprit au PDP, permet de prendre en compte à la fois les non linéarités et les interactions entre les variables. Ils appliquent cette méthode sur un algorithme de *Gradient Tree Boosting* et une base de prêts hypothécaires. Leurs résultats montrent que les facteurs déterminants du défaut sont sans surprise le ratio prêt / valeur et le taux d'intérêt. Mais leur méthode permet de mettre en évidence que l'influence de ces variables varie significativement suivant les caractéristiques des emprunteurs.

Causalité et prévision. Pour Grennepois et Robin (2019) ou Bracke et al. (2019), les méthodes permettant de mesurer et d'interpréter l'influence des prédicteurs sur le défaut ouvrent d'intéressantes pistes d'utilisation notamment en termes de modèles de stress tests. Toutefois, une telle utilisation des algorithmes de ML se situe aux marges de la prédiction (domaine naturel du ML) et de l'inférence causale (domaine traditionnel de l'économétrie). Pour bien comprendre la différence entre prévision et causalité, supposons que l'on dispose d'une base de données portant sur un ensemble de crédits pour lesquels on observe le type (bon / mauvais), un ensemble de caractéristiques des crédits et des emprunteurs, et un ensemble de variables macroéconomiques, comme par exemple le taux directeur de la Banque Centrale. Si l'on souhaite prévoir le type du crédit en fonction des caractéristiques individuelles et de l'environnement macroéconomique, il s'agit d'un problème de prévision qui se présente ici sous la forme d'une classification. En revanche, si l'on souhaite prévoir comment se comporteraient les crédits du portefeuille dans le cas d'une remontée des taux directeurs de 1%, il s'agit alors d'un problème d'inférence causale. Même si l'on observe par exemple une faible influence du taux directeur sur les défauts dans l'échantillon historique, rien ne permet d'affirmer qu'il en sera de même dans un nouvel environnement de hausse des taux. Il est impossible de répondre à cette question sans hypothèses supplémentaire ni structure. On touche ici aux limites du ML, même s'il existe aujourd'hui de nombreux travaux de recherche sur la notion de causalité en ML principalement dans le cadre de l'analyse des effets de traitement (voir par exemple Athey et Imbens (2019) pour la notion d'arbre de classification causal).

Instabilité et flexibilité. La flexibilité est l'un des grands atouts des algorithmes de ML car cette propriété leur confère la possibilité de détecter des non-linéarités susceptibles de mieux rendre compte du risque de crédit. Mais une trop grande flexibilité des modèles de scoring peut aussi s'avérer contre-productive si elle s'accompagne d'une perte de confiance vis-à-vis des résultats. Comment justifier que pour un même ensemble de prédicteurs, l'utilisation de tel ou tel algorithme de ML puisse conduire à des résultats très différents en termes de probabilité de défaut ? Si par ailleurs ces modèles ne sont pas facilement interprétables et re-

quièrent des techniques avancées comme le LIME pour expliquer les résultats, leur utilisation peut au final susciter des interrogations, voire des suspicions. Ainsi pour certaines applications, typiquement les modèles internes de risque utilisés pour le calcul des fonds propres réglementaires, les algorithmes de ML peuvent poser problème. Ainsi, dans ses recommandations sur l'estimation des paramètres de risque Bâlois, l'EBA (2017) pointe des écarts importants entre les paramètres de risque et les exigences de fonds propres de différentes institutions, qui ne reflètent pas les différences de profils de risque mais résultent plus de définitions différentes et de certains choix de modélisation. L'EBA conclut que la trop grande flexibilité de l'approche IRB est susceptible d'entrainer une perte de confiance vis-à-vis des modèles internes par les investisseurs et les autres acteurs du marché. Il est indéniable que l'utilisation de modèles de ML dans le contexte spécifique du calcul des fonds propres réglementaires ne ferait que renforcer ce problème.

2 L'apport des « Big » et des « New » Data

2.1 Nouvelles données et nouveaux acteurs

Lorsque l'on évoque l'apport du ML au scoring de crédit, il est difficile de distinguer les gains qui relèvent des algorithmes de ceux qui découlent de l'utilisation de *nouvelles* données. Mais qu'entend-on exactement par *nouvelles* données ? Ici le caractère novateur s'entend par opposition aux données traditionnelles généralement analysées dans les modèles de score de crédit. Dans le cas d'un crédit à la consommation, les variables explicatives incluent généralement des informations sur la nature du crédit, les caractéristiques de l'emprunteur (âge, revenus, situation matrimoniale), et sur son historique bancaire. L'exemple typique est le score FICO largement utilisé dans l'industrie financière américaine pour évaluer la solvabilité des clients particuliers. Ce score est construit à partir de différents facteurs tels que l'historique des paiements, la somme des encours de dette, la durée de l'historique de crédit, l'ouverture récente de nouveaux comptes, etc. A l'inverse, les nouvelles données proviennent quant à elles de sources beaucoup plus variées, souvent rendues accessibles par la digitalisation de la relation clientèle (données d'empreinte numérique) ou de données issues de nouvelles sources d'information sur les clients, telles que les réseaux sociaux. Il peut s'agir parfois de sources d'information très disparates, sans lien apparent avec la solvabilité des clients.

Les nouvelles données peuvent être collectées par les acteurs traditionnels (les banques) ou par des nouveaux acteurs (Fintechs[11]). On doit distinguer ici deux types d'utilisation de ces données suivant le type de Fintech considéré. L'utilisation la plus standard concerne les

[11] Dans cet article, le terme Fintech est utilisé pour désigner les entreprises qui utilisent les nouvelles technologies dans le secteur financier, et non les technologies elles-mêmes.

Fintechs qui s'apparentent à des établissements de crédit, comme par exemple les plateformes de crédit, les banques en ligne, les néo-banques, ainsi que certains sites marchands. Dans ce cas, les données servent directement à construire des scores qui déterminent en interne l'octroi du crédit, les conditions du financement, et/ou qui servent à contrôler les risques du portefeuille de prêts. La seconde utilisation de ces nouvelles données consiste à construire un score de risque de crédit qui est vendu à un établissement de crédit. Cette externalisation de la collecte et de l'exploitation des nouvelles données s'apparente donc au fonctionnement des scores traditionnels (FICO) mais, selon la nature des données collectées, elle soulève des enjeux spécifiques en termes de responsabilité juridique et de régulation.

Afin d'illustrer la « datadiversité » de l'industrie de scoring de crédit, nous présentons ci-dessous quelques exemples. Ainsi, la Fintech Big Data Scoring propose d'intégrer aux modèles de score d'octroi des banques des données de médias sociaux relatives à l'entreprise et/ou à son gérant, mais aussi des données relatives au mode de navigation (adresse IP, appareil utilisé, comportement de navigation, etc.) dans le cadre des demandes de prêts en ligne. La start-up NeoFinance utilise des données relatives à la qualité de l'emploi occupé par la personne qui sollicite le prêt et la qualité de ses connexions professionnel via le réseau LinkedIn. La Fintech Lenddo vise à développer l'inclusion financière dans les pays en voie de développement, en mobilisant des données non traditionnelles pour fournir à la fois une notation de crédit (*Lenddo Score*), mais aussi une vérification d'identité (*Lenddo Verification*). La stratégie de Lenddo est clairement de contourner le besoin d'un score de crédit officiel (de type FICO ou *Credit Bureau*) pour permettre au plus grand nombre d'avoir accès au crédit. Leur notation mobilise différentes sources d'information : l'activité du client sur les réseaux sociaux (Facebook, LinkedIn, Twitter, etc.), les connexions avec des personnes présentant de bons risques, les données de navigation issues des smartphones ou des ordinateurs sur lesquels s'effectuent la demande de prêt, des informations sur la manière dont la personne remplit le formulaire en ligne de la demande de crédit, etc. Enfin, on peut citer ZestFinance, une Fintech fondée par l'ancien *Chief Information Officer* de Google. Leur solution ZAML (*Zest Automated Machine Learning*) permet de construire un score à partir d'informations très disparates telles que le nombre de fois où le client a déménagé, l'utilisation ou non des lettres capitales dans le formulaire, etc., tout en conservant une interprétabilité du modèle et des décisions individuelles. Il est intéressant de noter que ZestFinance a débuté son activité en tant que préteur, puis a pivoté vers la construction et la vente de scores fondés sur l'intelligence artificielle et la mobilisation de nouvelles données.

Cette « datadiversité » est tout aussi grande du côté des plateformes de crédit qui ont accès à des informations privées qui ne sont pas utilisées ou qui ne sont pas disponibles,

pour des raisons légales, réglementaires ou organisationnelles, pour les prêteurs bancaires traditionnels. Selon Jagtiani et Lemieux (2019), ces sources d'informations peuvent porter sur l'historique des paiements des consommateurs (loyer, téléphone, pension alimentaire, etc.), les informations sur les flux de trésorerie des comptes bancaires (salaires, prélèvements, etc.), les transactions effectuées par carte de crédit, les remboursements pour dépenses de santé, les dépenses d'éducation, les habitudes d'achat, etc. Ces données sont soit achetées auprès de fournisseurs externes ou soit collectées avec l'assentiment des emprunteurs. Elles sont souvent complétées par des données d'empreinte numérique captées lors de la demande de prêt ou dans les historiques de navigation. Cela concerne typiquement les informations relatives au moment où est effectuée la demande de prêt (par exemple, effectuer une demande de prêt à 3 heures du matin peut ne pas être un bon signal) ou au lieu à partir de laquelle la demande est faite (par exemple, dans une zone de grande criminalité). Jagtiani et Lemieux (2019) montrent que la corrélation entre les scores propriétaires attribuées par la plateforme de crédit LendingClub et les scores FICO obtenus pour un échantillon de crédits comparables est passée d'environ 0.8 pour les prêts contractés en 2007 à moins de 0.35 pour les prêts contractés en 2015. Cette évolution traduit la richesse informationnelle de ces nouvelles sources de données par rapport aux données traditionnelles. Tang (2019) étudie la valeur économique que les emprunteurs attribuent à certaines informations confidentielles divulguées lors d'une demande de prêt. En exploitant une expérience contrôlée réalisée par une plateforme de crédit en ligne chinoise, elle estime qu'en moyenne un emprunteur exige une réduction d'environ 9% de la valeur actualisée nette de son prêt pour fournir en échange une information ultraconfidentielle telle que les coordonnées téléphoniques de son employeur.

Enfin, ces nouvelles données peuvent être mobilisées par d'autres acteurs commerciaux. Berg et al. (2019) considèrent ainsi le cas d'une importante société de E-commerce basée en Allemagne. Cette société permet à ses clients de ne régler leurs achats réalisés en ligne qu'à la réception de la marchandise sous un délai de 14 jours. Chaque transaction s'apparente à un crédit à la consommation de court terme, ce qui suppose que l'entreprise soit en mesure d'évaluer précisément la solvabilité de se clients. Pour cela, elle utilise les données d'empreinte numérique[12] laissée par les clients lors de leur achat en ligne, telles que le type d'appareil utilisé lors de la connexion (ordinateur, tablette, smartphone), le système d'exploitation (Windows, iOS, Android), le type d'adresse mail (Gmail, Hotmail, etc.), l'heure de connexion, l'adresse mail du client, etc.

L'apparition de ces nouveaux acteurs du marché du crédit, que ce soient des établisse-

[12] L'empreinte numérique désigne l'ensemble des traces laissées volontairement ou non par un utilisateur sur internet ou sur d'autres services informatiques (serveurs, service de sauvegarde).

ments de crédit, des entreprises marchandes ou des prestataires de services, pose la question de leur pouvoir de marché respectif et de l'accès à l'information source. Une crainte dans ce domaine est liée à l'intérêt que pourraient porter au marché du crédit les géants de la technologie (ou BigTech) tels que les GAFA, Tencent ou Alibaba. Ces sociétés ont l'avantage d'avoir accès aux informations sur les activités commerciales de leurs clients en tant que sous-produit de leurs services de commerce électronique, de paiement, ou de réseaux sociaux (voir par exemple Frost et al. (2019)). Ces grandes entreprises de technologie peuvent observer la tendance des ventes et les flux de trésorerie des entreprises actives dans leur espace de commerce électronique, réduisant ainsi considérablement le coût de la collecte d'informations. En ayant un accès privilégié et gratuit à ces nouvelles sources d'informations, ces BigTechs pourraient être en mesure de mieux évaluer la solvabilité des emprunteurs, d'offrir des conditions tarifaires plus avantageuses que la concurrence et d'exiger de moindres garanties que les banques traditionnelles.

2.2 Nouvelles données versus données massives

Il existe une confusion entre les notions de *nouvelles données* (*New Data*) et de *méga-données* (*Big Data*). Pour bien comprendre la différence, il convient de revenir sur la notion de Big Data. Pour ce faire, imaginons un fichier Excel de données dans lequel les p colonnes représentent des variables (prédicteurs) et les n lignes des individus (instance). On parle de *Big Data* dans trois configurations distinctes qui n'ont pas les mêmes implications sur le plan de la modélisation statistique (pour plus de détails voir Varian (2014) ou Charpentier, Flachaire, et Ly (2018)). La première configuration, que Varian (2014) qualifie de *Tall Data* (données de grande taille), correspond à une situation où l'on dispose d'observations pour un très grand nombre d'individus. La dimension n est alors largement supérieure au nombre de prédicteurs p. Cette configuration ne pose pas de problème particulier sur le plan statistique. La seule conséquence est que la convergence des estimateurs rend généralement caduque le besoin d'inférence sur les paramètres des modèles.

La seconde configuration, que Varian (2014) qualifie de *Fat Data* (données épaisses), correspond à une situation où le nombre de prédicteurs p est très élevé relativement au nombre d'instances n. C'est typiquement le cas lorsque l'on considère des données de réseaux sociaux puisqu'on observe pour chaque individu une multitude de connections, chacune représentée par une variable. Une telle configuration, que l'on qualifie parfois de malédiction de la dimensionalité, pose le problème du surajustement (*overfitting*). Admettons que l'on construise un modèle de régression avec l'ensemble de ces variables et que l'on puisse estimer ses paramètres, la qualité de l'ajustement sur l'échantillon d'estimation sera largement plus grande que celle

que l'on pourra obtenir hors-échantillon d'estimation (*out-of-sample*). La solution réside alors dans la sélection des variables, mais les méthodes usuelles de l'économétrie posent alors le problème de la multiplicité des tests à mettre en œuvre, du contrôle du niveau de risque de ces tests multiples, et du temps de calcul. C'est donc précisément dans cette configuration que les méthodes de ML trouvent tout leur intérêt. L'exemple typique est celui des méthodes de régressions pénalisées (*Lasso*, *Ridge*, etc.) qui permettent de sélectionner les variables pertinentes en pénalisant la valeur des paramètres de la régression de sorte à ce que les coefficients des variables non pertinentes soient les plus proches possibles de 0. De façon générale, les algorithmes de ML permettent d'exploiter des *Big Data* (au sens de *Fat Data*) qui n'auraient pas pu être exploitées autrement. Leur principal atout réside dans leur capacité à utiliser les données de l'échantillon d'apprentissage pour sélectionner de façon autonome une forme fonctionnelle optimale du lien entre la cible et les prédicteurs, alors qu'il aurait été impossible ou trop couteux de la déterminer manuellement parmi les p prédicteurs possibles.

Enfin, la troisième configuration correspond à des données massives, au sens où elles occupent trop de place en mémoire pour pouvoir être chargées et maniées dans des logiciels standards. Il s'agit ici avant tout d'un problème informatique qui peut être résolu de différentes manières, comme par exemple avec l'environnement *MapReduce / Hadoop*.

Ainsi, suivant cette définition, les nouvelles données ne sont pas toujours nécessairement de type *Big Data*. Par exemple, dans leur étude Berg et al. (2019) considèrent uniquement 10 variables d'empreinte numérique. A l'inverse, cela sera le cas pour des données collectées sur réseaux sociaux ou pour des données concernant les délais de paiement des entreprises vis-à-vis de leurs fournisseurs. De même, les nouvelles données ne nécessitent pas toujours l'utilisation d'algorithmes de ML. Berg et al. (2019) considèrent une simple régression logistique pour construire leur modèle de scoring à partir de l'empreinte numérique des clients. A l'inverse, les Fintechs spécialisées dans le scoring qui utilisent généralement plusieurs centaines de prédicteurs sont contraintes d'adopter des techniques de ML.

2.3 Les nouvelles données améliorent-elles le scoring de crédit ?

Gains globaux de performances. Sur le plan théorique, Wei et al. (2016) évaluent l'impact des données de réseaux sociaux sur la qualité des modèles de scoring de crédit et sur la formation endogène des liens entre les consommateurs. Les auteurs montrent que les consommateurs qui souhaiteraient améliorer leur score seront amenés à former moins de liens dans leurs réseaux sociaux et à privilégier des partenaires similaires. L'impact de ces comportements stratégiques sur la performance prédictive des modèles de score est alors ambigu. D'un côté, l'usage des réseaux sociaux tend à améliorer la précision des scores puisque l'on

augmente l'ensemble d'information en mobilisant des informations sur des individus similaires à l'emprunteur. Toutefois, cette utilisation peut engendrer une plus grande fragmentation de ces réseaux, qui tend à réduire les gains prédictifs. Sur le plan empirique, l'apport des données de réseaux sociaux est très discuté. Plusieurs Fintechs (Lenddo, Big Data Scoring, etc.) vantent l'intérêt de ces données, qui constituent par ailleurs leur cœur de métier. D'autres, comme ZestFinance, rejettent l'utilisation des données de réseaux sociaux en mettant en cause à la fois leur utilité et la légitimité de leur utilisation.

Berg et al. (2019) montrent que la prise en compte de l'empreinte numérique des clients permet d'améliorer sensiblement les performances prédictives des modèles de score d'une entreprise de E-commerce. L'AUC de leur modèle de régression logistique basé sur les seules variables d'empreinte atteint 69.6% contre 68.3% pour un modèle basé sur un score de crédit de type FICO. De plus, la combinaison des scores de crédit de type FICO et des variables d'empreinte numérique permet d'atteindre une AUC de 73,6%, soit 5,3% de plus que l'AUC du modèle fondé uniquement sur les scores de crédit. Ce gain, obtenu à partir d'une simple régression logistique, confirme clairement que les données d'empreinte numérique apportent une information complémentaire par rapport aux données traditionnelles mobilisés dans les modèles de score bancaire.

Plus récemment, Óskarsdóttir et al. (2019) modélisent le défaut de titulaires de carte de crédit en utilisant des relevés détaillés de téléphonie mobile permettant de reconstruire le réseau social du titulaire. Les données anonymisées portent sur 90 millions de numéros de téléphone, 2 millions de clients bancaires et incluent différentes informations sociodémographiques et bancaires. Le réseau de chaque individu est synthétisé par environ 200 statistiques (Page Rank, nombre de liens, etc.) décrivant les liens du client avec d'autres clients ayant eu des retards de paiements dans le passé ou d'autres incidents bancaires. Les auteurs montrent que l'inclusion des caractéristiques du réseau téléphonique permet d'augmenter l'AUC du modèle de score (régression logistique, arbre de classification ou Random Forest), comparativement à l'AUC obtenue uniquement sur les données sociodémographiques et/ou les données bancaires. Ils montrent en outre que l'AUC d'un modèle construit uniquement avec les données de ce réseau est comparable à celle obtenue avec les données bancaires, ce qui ouvre des perspectives intéressantes notamment pour les pays en voie de développement, peu bancarisés mais dans lesquels l'usage du portable est généralisé.

Des résultats similaires ont été obtenus par Frost et al (2019) dans leur analyse de la plateforme argentine Mercado Libre, spécialisée dans la distribution de crédits aux petites entreprises. Les auteurs montrent que les techniques d'évaluation du crédit basées sur le ML surperforment les notations du bureau de crédit en termes de prévision des taux de

perte, notamment pour les firmes les plus risquées. Ils comparent notamment les courbes ROC obtenues à partir d'une régression logistique basée sur les seules notations du bureau de crédit, d'une régression logistique basée sur le score interne de Mercado (lui-même fondé sur un ensemble d'informations disparate concernant l'entreprise), et d'une méthode de ML utilisant uniquement le score interne de Mercado.

Nouvelles données et inclusion financière. Est-ce que le recours à ces nouvelles données permet à des individus ou des entreprises d'accéder à des crédits alors qu'ils en auraient été exclus sur la base d'informations plus traditionnelles ? Plusieurs raisons laissent à penser que les nouvelles données couplées au traitement automatique, peuvent contribuer à améliorer l'accès au crédit. Bazarbash (2019) avance ainsi que les banques traditionnelles s'abstiennent souvent d'évaluer le risque de crédit des petits emprunteurs car la faible espérance de gain et le risque potentiellement élevé du prêt ne permettent pas de couvrir les coûts. En automatisant le processus de notation et en utilisant ces nouvelles sources de données, les FinTechs peuvent mieux évaluer la solvabilité des petits emprunteurs plus risqués en contractant fréquemment de petits montants de prêts et en surveillant leur comportement de remboursement. Ainsi, les FinTech peuvent garantir un meilleur accès au crédit, typiquement pour des petites entreprises ne disposant pas des garanties financières requises dans le secteur traditionnel. Schweitzer et Barkley (2017) analysent une importante base de données de crédits contractés par les petites entreprises américaines en 2015 (*Federal Reserve's Small Business Credit Survey*). Leur étude montre que les entreprises bénéficiant d'un financement via des plates-formes en ligne partagent les mêmes caractéristiques que celles à qui les banques traditionnelles ont refusé un crédit. Ce résultat suggère que les FinTechs contribuent à l'inclusion financière de ces petits emprunteurs, mais qu'il reste difficile de relier ce résultat à des sources de données en particulier.

Jagtiani et Lemieux (2019) montrent que les sources alternatives de données utilisées par LendingClub permettent à certains clients qui auraient été classés dans les catégories les plus risquées selon les critères traditionnels, de devenir de « bons » risques et ainsi d'obtenir des prêts à de meilleures conditions. Environ 8% des emprunteurs notés A (meilleure note) par la méthode de scoring de LendingClub avaient obtenu des scores FICO inférieurs à 680 (ratings « *poor* » ou « *fair* »), et 28% des emprunteurs notés B avaient des scores FICO se situant dans la même plage. De même, Berg et al. (2018) trouvent que les données d'empreinte numérique permettent d'accepter des emprunteurs au profil fragile, mais avec un bon score d'empreinte numérique, qui n'auraient pas été acceptés uniquement sur la base des cotes de crédit. A l'opposé, ces données conduisent aussi à rejeter des demandes émanant d'emprunteurs qui

auraient été acceptés sur la base uniquement de leur score de crédit officiel.

2.4 Les risques liés aux nouvelles données de scoring

Enjeux juridiques et éthiques. Le recours à de nouvelles données est une source de tension entre d'une part la volonté des banques et des régulateurs de mesurer le risque avec précision et d'autre part la protection des données personnelles des clients. De façon caricaturale, nous pourrions dire que Bâle III/IFRS9 est en conflit avec le Règlement Général sur la Protection des Données (RGPD). En Europe, le RGPD constitue en effet le texte de référence en matière de protection des données à caractère personnel. Ce règlement repose sur plusieurs principes parmi lesquels les principes de sécurité et de confidentialité des données, de finalité, de proportionnalité et de pertinence, et finalement le principe d'une durée de conservation limitée. Dans le cas du scoring de crédit, ce règlement n'interdit donc en rien l'utilisation de la plupart des nouvelles sources de données actuellement mobilisées. Il existe en effet plusieurs techniques d'anonymisation des données, souvent utilisées en médecine, qui permettent de partager avec des tiers des données personnelles, par exemple d'empreinte numérique ou de réseaux téléphonique. Un tel partage peut s'opérer aujourd'hui de façon parfaitement sécurisée. Cependant, dans le cas de l'externalisation de la construction du score par une Fintech, la banque en tant qu'ordonnateur du traitement engage sa propre responsabilité vis-à-vis des manquements à ces protocoles. De la même façon, comment s'assurer que les données transmises par la banque à la Fintech pour construire le modèle de score ne sont pas conservées au-delà de cette phase de développement ?

Au-delà de ces enjeux juridiques, des questionnements éthiques se posent quant à l'utilisation de certaines de ces nouvelles données, notamment des données de réseaux. Dégrader les conditions d'accès au crédit d'un client particulier, toutes choses égales par ailleurs, parce qu'il côtoie de mauvais payeurs n'est pas juridiquement condamnable, mais cela pose clairement problème. A l'inverse, les mêmes sources de données peuvent favoriser l'inclusion financière d'autres clients qui côtoient des bons risques. Dès lors, comment arbitrer sur l'acceptabilité sociale de ces données ? Óskarsdóttir et al. (2019) proposent une solution inspirée des grilles de score. Dans le domaine du scoring, il est courant de discrétiser les variables explicatives continues et d'attribuer un score à chacun des segments en fonction de leur contribution à la détection du défaut. Une utilisation éthique de ce principe consisterait à attribuer un score nul aux segments qui désavantageraient les emprunteurs, tout en laissant des poids positifs aux segments qui faciliteraient l'accès au crédit. Appliquer une telle pénalisation éthique aux variables issues des nouvelles sources d'information, conduirait certes à dégrader les performances prédictives du modèle, mais garantirait le caractère socialement acceptable de l'utilisation de ces

nouvelles données.

Toutefois, le critère ultime qui permettra de lever cette ambiguïté au sujet de l'utilisation des nouvelles données, est celui de l'acceptabilité par les clients. Au-delà de la question morale ou juridique, la question est de savoir si les clients, particuliers ou entreprises, sont prêts à voir certaines de leurs données personnelles utilisées dans le cadre de leur demande de prêt. Prenons l'exemple des crédits immobiliers : il est connu qu'en dehors du cycle économique et du taux de chômage, un des prédicteurs les plus importants du défaut est le divorce. Dès lors, toute variable permettant de prévoir le divorce sera un bon prédicteur du défaut. Si une Fintech venait à utiliser un score fondé sur l'analyse de consultation de sites de rencontres extra-conjugales, est-ce que les clients seraient prêts à accepter ce type de démarche dans le but d'obtenir des conditions de prêt plus avantageuses ? Au-delà même des aspects juridiques ou éthiques, la technologie n'a de valeur que dans la mesure où cette technologie est acceptée au final par le client. Cela sera sans aucun doute le principal frein au développement de l'intelligence artificielle dans le domaine du risque de crédit, comme dans beaucoup d'autres domaines.

New Data, ML et biais. Le principal risque de l'association entre le ML et les nouvelles sources de données réside dans l'apparition de biais ou de traitements inéquitables (ACPR, 2018). Appliquée au domaine du scoring de crédit, la question revient à déterminer si les algorithmes de ML peuvent aboutir à pénaliser certaines populations, voire à les exclure totalement de l'accès au crédit, sans même que cela soit voulu par l'institution financière.

La question des biais s'est posée dès que les banques ont commencé à utiliser des approches statistiques pour déterminer les conditions d'octroi des crédits. Ainsi, dès la fin des années 70, Hsia (1978) propose une réflexion juridique sur les méthodes de scoring et s'interroge sur leur compatibilité avec la loi *Equal Credit Opportunity Act* (ECOA) votée par le congrès américain en 1976, et dont l'objectif était précisément de garantir un accès équitable au crédit. On peut citer également les travaux de Chandler et Ewert (1976) qui s'intéressaient quant à eux à la question de la discrimination sexuelle dans l'accès au crédit.

De nos jours, comment ces biais peuvent-ils émerger dans le contexte du ML et du Big Data ? Tout d'abord, l'algorithme de ML peut tout simplement sélectionner parmi l'ensemble des prédicteurs disponibles, des variables considérées comme discriminatoires telles que le genre, l'origine ethnique, l'orientation sexuelle ou politique, etc. Par définition, la machine ne sait pas ce qui est moral ou pas, ce qui est juridiquement autorisé ou non. Elle ne peut donc pas définir les notions de discrimination ethnique, de genre, ou de religion. C'est là que réside le principal danger de l'apprentissage « machine » par rapport à l'apprentissage «

humain ». Un modélisateur humain au sein d'une institution financière, ne prendra pas le risque moral, juridique ou de réputation d'inclure ce type de variable dans son modèle de score, quand bien même les données seraient disponibles. A l'inverse, les algorithmes de ML sélectionnent les variables pertinentes de façon autonome, par minimisation d'un critère sur un échantillon d'apprentissage, et sans intervention humaine. Ainsi rien ne garantit que les algorithmes ne sélectionnent pas de variables discriminatoires. Le risque est d'autant plus grand que l'on considère des algorithmes non interprétables qui s'apparentent à des boites noires. Dans ce cas, il est indispensable de mettre en œuvre des méthodes de type *Model-Agnostic Methods* permettant de rendre les modèles interprétables *ex-post*, et de détecter in fine l'éventuelle influence de variables discriminatoires. Une façon beaucoup plus simple de régler le problème consiste à vérifier la conformité des données sources. Rappelons qu'en France, le simple fait de collecter et de détenir certaines informations jugées discriminatoires engage la responsabilité de l'institution de crédit. Cependant, comme nous allons le voir ci-dessous, l'absence de variables discriminatoires dans les données sources ne garantit en rien l'absence de biais dans les modèles de scoring.

En effet, les biais peuvent apparaitre plus subtilement de façon indirecte, au travers de variables dites proxy. On parle alors de proxy discrimination (Prince et Schwarcz, 2019). L'idée générale est que la discrimination résulte de l'interaction ou triangulation de plusieurs variables qui n'apparaissent pas en soi discriminatoires. Par exemple, un algorithme de ML peut croiser plusieurs variables licites telles que le revenu et le type de logement afin de prévoir implicitement le lieu de résidence, et utiliser cette information pour discriminer des clients résidant dans des zones géographiques sensibles. Le risque est d'autant plus grand que les bases de données comportent de nombreux prédicteurs (*Fat Data*), puisque les algorithmes de ML peuvent alors identifier des interactions entre un grand nombre de variables.

Fuster et al. (2018b) montrent que l'utilisation d'un modèle de ML, plus flexible et plus performant au niveau global qu'une approche paramétrique standard, génère par définition des prévisions plus dispersées. L'adoption d'une telle technologie crée nécessairement des gagnants et des perdants par rapport à la situation de référence, c'est-à-dire des individus classés initialement comme risqués qui deviennent non risqués et vice-versa. A partir d'une base de prêts hypothécaires aux Etats-Unis, les auteurs montrent que dans le cas du passage d'un scoring par régression logistique à une approche de ML, les emprunteurs noirs et hispaniques sont globalement perdants, tandis que les emprunteurs blancs sont globalement gagnants. Toutefois, pour Fuster et al. (2018b), ces disparités sont principalement attribuables à la flexibilité accrue de la méthode de classification.

A l'inverse, Bartlet et al. (2019) montrent que le ML permet de réduire les discriminations

ethniques[13] sur le marché américain du crédit immobilier. Ils croisent pour cela différentes bases de données, dont le *Home Mortgage Disclosure Act* qui couvre près de 90% des crédit hypothécaires aux Etats-Unis sur la période 2009-2015. Leurs résultats montrent que les prêteurs traditionnels facturent respectivement 7,9 et 3,6 points de base de plus aux emprunteurs latino-américains et afro-américains, toutes choses égales par ailleurs. Au niveau agrégé, cela représente près de 765 millions de dollars par an en intérêts supplémentaires. Or, si les FinTech discriminent également, la discrimination est dans ce cas 40% moindre que celle des prêteurs traditionnels. De même, ils observent que les prêteurs traditionnels rejettent les demandes des latino-américains et afro-américains environ 6% plus souvent qu'ils ne rejettent les demandes de clients qui ne sont pas issus de ces minorités. A l'échelle macroéconomique sur la période 2009-2015, cela représente de 0,74 à 1,3 million de clients latino-américains et afro-américains dont les crédits auraient pu être acceptés s'il n'y avait pas eu de discrimination.

La question des biais est devenue un argument commercial pour certaines FinTech. Ainsi ZestFinance propose à ses clients bancaires de vérifier ex-post si le score produit par ses algorithmes de ML conduit à des biais vis-à-vis de certaines populations. La banque peut alors augmenter ou diminuer l'influence de certaines variables dans le modèle de façon à réduire ces biais, tout en contrôlant la validité des prévisions de défaut. ZestFinance affirme ainsi que si son outil de scoring était appliqué partout aux Etats-Unis, cela permettrait de réduire de 70% l'écart des taux d'approbation sur les crédits hypothécaires entre les emprunteurs blancs et hispaniques, et de 40% l'écart avec les emprunteurs afro-américains, permettant à plus de 172 000 personnes chaque année de devenir propriétaire.

3 Conclusion

Les approches traditionnelles du scoring de crédit alliant différents prétraitements sur les données et des approches paramétriques simples, telles que la régression logistique, offrent de très bonnes performances. Dès lors, lorsque l'on raisonne à ensemble d'information constant, les algorithmes de ML, y compris les plus récents, n'apportent que des gains de performance marginaux, même s'ils permettent parfois des gains de productivité. En revanche, ces techniques permettent de mobiliser de nouvelles sources d'information, qui autrement n'auraient pas pu être intégrées dans les modèles de score, en raison de leur grande dimension. L'alliance du ML et des nouvelles données, parfois sous la forme de Big Data, permet alors de révéler des signaux faibles que ce soit sous la forme d'interactions ou de non-linéarités, qui sans que l'on sache toujours les expliquer, améliorent l'évaluation de la solvabilité des clients. Plus fon-

[13] Rappelons qu'aux Etats-Unis, les informations sur l'origine ethnique peuvent être collectées et utilisées sous certaines conditions.

damentalement, ces gains prédictifs globaux se déclinent parfois au niveau microéconomique par des gains individuels, en améliorant notamment l'inclusion financière et l'accès au crédit des emprunteurs les plus fragiles. Ainsi, la recherche dans le domaine de la modélisation du risque de crédit ne doit pas tant tendre à développer de nouvelles approches méthodologiques de classification, que de tirer parti de nouvelles sources de données innovantes. Toutefois ces nouvelles sources de données soulèvent de nombreuses questions éthiques, juridiques et réglementaires. Couplées à des techniques de ML difficilement interprétables ou mal maitrisées, l'utilisation de ces données peut engendrer des biais et conduire à dégrader les conditions d'accès au crédit de populations entières sur la base de critères ethniques, religieux, sociaux, etc. sans même que l'établissement de crédit ou la Fintech ne s'en aperçoive. Ces opportunités, mais aussi ces risques, appellent sans aucun doute la mise en œuvre d'une nouvelle forme de régulation financière basée sur une certification des algorithmes et des données mobilisées par les établissements de crédit.

Références

[1] Altman, E., Marco, G., et Varetto, F., (1994), Corporate distress diagnosis : Comparisons using linear discriminant analysis and neural networks (the Italian experience), Journal of Banking and Finance 18, 505-529.

[2] ACPR, (2018), Intelligence artificielle : enjeux pour le secteur financier, Document de réflexion, Décembre 2018.

[3] Athey, S., (2019), The impact of machine learning on economics, in The economics of artificial intelligence : an agenda, Ajay Agrawal, Joshua Gans, et Avi Goldfarb, ed., 507-547.

[4] Athey, S., et Imbens, G.W., (2019), Machine Learning methods that economists should know about Annual Review of Economics, 11, 685-725.

[5] Baesens, B, Van Gestel, T, Viaene, S, Stepanova, M, Suykens, J, et Vanthienen, J. (2003), Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society, 54(6), 627-635.[2]

[6] Bartlett, R., Morse, A., Stanton, R., et Wallace, N., (2019), Consumer-lending discrimination in the fintech era, NBER working paper 25943.

[7] Bazarbash, M., (2019), FinTech in financial inclusion Machine Learning applications in assessing credit risk, IMF working paper 19-109.

[8] Berg, T., Burg, V., Gombović, A., et Puri, M., (2019), On the rise of FinTechs : Credit scoring using digital footprints. Michael J. Brennan Irish Finance Working Paper Series Research Paper No. 18-12.

[9] Bracke, P., Datta, A., Jung, C., et Sen, S., (2019), Machine learning explainability in finance : an application to default risk analysis, Bank of England, Staff Working Paper No. 816.

[10] Candelon, B., Dumitrescu, E., et Hurlin, C., (2012), How to evaluate an Early Warning System ?, IMF Economic Review, 60(1), 75-113.

[11] Carter, C., et Catlett, J., (1987), Assessing credit card applications using Machine Learning. IEEE Expert 2, 71-79.

[12] Chandler, G. G., et Ewert, D. C., (1976), Discrimination on basis of sex and the Equal Credit Opportunity Act, Credit Research Centre, Purdue University, Indiana.

[13] Charpentier, A., Flachaire, E., et Ly, A. (2018), Econometrics and Machine Learning, Economics and Statistics 505-506, 147-169.

[14] Coffman, J. Y., (1986), The proper role of tree analysis in forecasting the risk behaviour of borrowers, Management Decision Systems, Atlanta, MDS Reports.

[15] Desai, V. S., Crook, J. N., et Overstreet, G. A., (1996), A comparison of neural networks and linear scoring models in the credit environment, European Journal of Operational Research, 95, 24-37.

[16] EBA (2016), Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013, September 2016.

[17] EBA (2017), Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures.

[18] Fuster, A., Plosser, M., Schnabl, P., et Vickery, J., (2018a), The role of technology in mortgage lending, NBER Working Paper No. 24500.

[19] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., et Walther, A., (2018b), Predictably unequal ? The effects of Machine Learning on credit markets ? Working Paper SSRN.

[20] Frost, J., Gambacorta, L., Huang, Y., Shin, H.S., et Zbinden, P., (2019), BigTech and the changing structure of financial intermediation, BIS Working Paper No. 779.

[21] Grennepois, N., Alvirescu, M.A., et Bombail, M. (2018), Using Random Forest for credit risk models, Deloitte Risk Advisory, Septembre 2018.

[22] Grennepois, N. et Robin, E., (2019), Explain artificial intelligence for credit risk management, Deloitte Risk Advisory, Juillet 2019.

[23] Hisa, D.C., (1978), Credit scoring and the Equal Credit Opportunity Act, Hasting Law Journal, 30(2), 371-448.

[24] Jagtiani, J., et Lemieux, C., (2019), The roles of alternative data and Machine Learning in Fintech lending : evidence from the LendingClub consumer platform, WP18-15, FRB of Philadelphia.

[25] Lessmann, S., Baesens, B., Seow, H.V., et Thomas, L.C., (2015), Benchmarking state-of-the-art classification algorithms for credit scoring : An update of research, European Journal of Operational Research, 247(1), 124-136.

[26] Loterman, G., Brown, I., Martens, D., Mues, C., et Baesens, B., (2012), Benchmarking regression algorithms for loss given default modeling, International Journal of Forecasting, 28(1), 161-170.

[27] Makowski, P., (1985), Credit scoring branches out, The Credit World, 75, 30-37.

[28] Molnar, C., (2019), Interpretable Machine Learning, A Guide for Making Black Box Models Explainable. https ://christophm.github.io/interpretable-ml-book/

[29] Mullainathan, S., et Spiess, J. (2017), Machine Learning : An applied econometric approach, Journal of Economic Perspectives, 31 (2), 87-106.

[30] Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., et Baesens, B., (2019), The value of big data for credit scoring : Enhancing financial inclusion using mobile phone data and social network analytics, Applied Soft Computing, 74, 26-39.

[31] Phaure H. et Sartre J. (2019), Classification non supervisée : utilisations innovantes en banque, Deloitte Risk Advisory, Avril 2019.

[32] Prince, A. et Schwarcz, D.B., (2019), Proxy discrimination in the age of artificial intelligence and big data, Iowa Law Review, forthcoming. Available at SSRN : https ://ssrn.com/abstract=3347959.

[33] Schweitzer, M.E., et Barkley, B., (2017), Is Fintech good for small business borrowers ? Impacts on firm growth and customer satisfaction, FRB of Cleveland working paper No. 17-01.

[34] Srinivasan, V., et Kim, Y. H., (1987), Credit granting : a comparative analysis of classification procedures. Journal of Finance, 42, 665-683.

[35] Tam, K.Y. et Kiang, M.Y. (1992), Managerial applications of neural networks : The case of bank failure predictions, Management Science, 38, 926-947.

[36] Tang, H., (2019), The value of privacy : Evidence from online borrowers, Working Paper, HEC Paris.

[37] Thomas, L.C., (2000), A survey of credit and behavioural scoring : forecasting financial risk of lending to customers. International Journal of Forecasting, 16, 149-172.

[38] Varian, H.R., (2014), Big data : New tricks for econometrics, Journal of Economic Perspectives, 28, 3-28.

[39] Wei, Y., Yildirim, P., Van den Bulte, C., et Dellarocas, C., (2016), Credit scoring with social network data, Marketing Science, 35 :2, 234-258.